ORIGINAL PAPER

Nagoya J. Med. Sci. **65**. 127 ~ 132, 2002

STATISTICAL METHOD IN A COMPARATIVE STUDY IN WHICH THE STANDARD TREATMENT IS SUPERIOR TO OTHERS

MITSURU IKEDA[1], KAZUHIRO SHIMAMOTO[2], TAKEO ISHIGAKI[3] and KAZUNOBO YAMAUCHI[1]

[1]Department of Medical Information and Medical Records, Nagoya University Hospital, 65, Tsurumai-cho, Showa-ku, Nagoya 466-8560, Japan.
[2]Division of Medical Radiological Technology, Department of Radiological Technology, Nagoya University School of Health Sciences, 1-1-20 Daikominami, Higashi-ku, Nagoya 461-8673, Japan.
[3]Department of Radiology, Nagoya University School of Medicine. 65, Tsurumai-cho, Showa-ku, Nagoya 466-8560, Japan.

ABSTRACT

The statistical method in a comparative study in which the standard treatment is theoretically or practically superior to the others has been investigated for a matched-pairs design. We derived this statistical method from one based on the maximum likelihood and score methods. Here we have shown that the score test statistic is algebraically the same as the statistic from the maximum likelihood method. As an example of our method's applications, we have considered a study on the detection of nodules on chest X-ray images displayed on a CRT with low luminance.

Key Words: equivalence tests, individually matched pairs, maximum likelihood method, score method.

INTRODUCTION

The luminance of electronic displays used over the long term will be so dark that radiologists will not accept images from such display devices in clinical situations. However, there is no subjective criterion for judging whether or not the luminance of electronic displays is acceptable for clinical diagnosis. Thus, to clarify the minimum level of luminance of electronic displays below which softcopy images cannot be used for medical image interpretation, we conducted an image-reading experiment in which the observers had to detect nodules in chest X-ray soft-copy images on a cathode ray tube (CRT) with various levels of luminance. In these experiments, for each image, the observation order was always from the darkest display condition to the brightest. Therefore, the detection rate of nodules on the CRT is always higher under brighter rather than darker luminance. From these results, by using the statistical method, we tried to determine the CRT monitor luminance level below which the detection rates are significantly inferior to those on a CRT monitor with standard luminance. What statistical

Address for correspondence: Mitsuru Ikeda, Department of Medical Information and Medical Records, Nagoya University Hospital, 65, Tsurumai-cho, Showa-ku, Nagoya 466-8560, Japan.
Phone: +81-52-744-2666; Fax: +81-52-744-2973
Email: a40495a@nucc.cc.nagoya-u.ac.jp

method should we use for addressing this problem?

Generally speaking, the statistical test for the above problem should examine whether or not the detection rate of signals in one considered treatment is reliably inferior to that in the standard treatment that is either theoretically or practically superior to any other considered treatment in the detection of signals. For such statistical testing, the experimental results will usually be summarized numerically as the number of signals detected both in a considered treatment and in the standard treatment, x_{11}, the number of signals undetected either in a considered treatment or in the standard treatment, x_{00}, the number of signals undetected in a considered treatment and those detected in the standard treatment, x_{01}, the number of signals detected in a considered treatment and those undetected in the standard treatment, x_{10}. In our experimental design, the same signals are used both in a considered treatment and in the standard treatment. Thus, in this comparative study, the samples can be considered as individually matched. In addition, it is central to this statistical method for the above problem that $x_{10} = 0$.

As far as we know, there is no statistical method that is specific to this type of problem. In this paper, we have investigated the statistical technique for addressing this type of problem, that is, the statistical method in the comparative study using a matched-pairs design in which the standard treatment is theoretically or practically superior to the others.

NOTATION

Let us denote the total number of signals that will be detected in the comparative study of two treatments as N. The symbols x_{11}, x_{00}, x_{01}, and x_{10} denote the same observed numbers as in the previous section. Here, $x_{10} = 0$. We now consider a probability model corresponding to the above observed numbers, and we denote the probability of detecting signals both in a considered treatment and in the standard treatment as p_{11}, the probability of detecting signals neither in a considered treatment nor in the standard treatment as p_{00}, and the probability of detecting signals not in a considered treatment but in the standard treatment as p_{01}. Thus, the observation numbers and probabilities in this problem are classified in Table 1. Here, the following equations hold:

$$x_{11} + x_{01} + x_{00} = N, \tag{1}$$

$$p_{11} + p_{01} + p_{00} = 1. \tag{2}$$

Table 1. Observations and probability model

Observations			Probability model		
	Standard treatment			Standard treatment	
Considered treatment	Detected	Not detected	Considered treatment	Detected	Not detected
Detected	x_{11}	0	Detected	p_{11}	0
Not detected	x_{01}	x_{00}	Not detected	p_{01}	p_{00}

TEST STATISTICS

Maximum likelihood method

The log likelihood for the observations and the probability model given in the previous section is expressed as

$$L = x_{01}\ln p_{01} + x_{11}\ln p_{11} + (N - x_{01} - x_{11}) \ln (1 - p_{01} - p_{11}) \tag{3}$$

Let us denote the maximum likelihood estimator (MLE) of p_{01} and p_{11} as $\hat{p}_{01}$ and $\hat{p}_{11}$ respectively. From the above likelihood, these are trivial, and are given as,

$$\hat{p}_{01} = x_{01} / N, \tag{4}$$

$$\hat{p}_{11} = x_{11} / N. \tag{5}$$

Here, let us denote the Fisher information matrix for p_{01} and p_{11} as **I**, and denote the inverse matrix of **I** as $\mathbf{I}^{-1}$. From Eq. (3), **I** is given as

$$\mathbf{I} = \begin{pmatrix} \dfrac{N(1-p_{11})}{p_{01}(1-p_{01}-p_{11})} & \dfrac{N}{1-p_{01}-p_{11}} \\ \dfrac{N}{1-p_{01}-p_{11}} & \dfrac{N(1-p_{01})}{p_{11}(1-p_{01}-p_{11})} \end{pmatrix}, \tag{6}$$

and $\mathbf{I}^{-1}$ is given as

$$\mathbf{I}^{-1} = \begin{pmatrix} \dfrac{p_{01}(1-p_{01})}{N} & -\dfrac{p_{01}p_{11}}{N} \\ -\dfrac{p_{01}p_{11}}{N} & \dfrac{p_{11}(1-p_{11})}{N} \end{pmatrix}, \tag{7}$$

Next, we would like to establish that the detection rate of nodules in the considered treatment is reliably inferior to the one in the standard treatment. Here, let us consider a sufficiently small number, δ, and consider a one-sided test for the null hypothesis H_0: $p_{01} = \delta$ against the alternative H_1: $p_{01} > \delta$. The vector consisting of MLE, $\hat{\mathbf{p}} = (\hat{p}_{01}, \hat{p}_{11})^t$ is known to have the following asymptotic properties; with a large N, $\hat{\mathbf{p}}$ will be approximately multi-normally distributed with the means of $(p_{01}, p_{11})^t$ and the variance-covariance matrix of $\mathbf{I}^{-1}$. The MLE of p_{11} for a given value of $p_{01} = \delta$, $(\hat{p}_{11})_{p_{01}=\delta}$, is

$$(\hat{p}_{11})_{p_{01}=\delta} = \frac{x_{11}(1-p_{01})}{N - x_{01}} . \tag{8}$$

By using this MLE of p_{11} under $p_{01} = \delta$, $(\hat{p}_{11})_{p_{01}=\delta}$, and substituting p_{01} with δ, under the null hypothesis H_0, $\mathbf{I}^{-1}$ can be approximated as

$$\hat{\mathbf{I}}^{-1} = \begin{pmatrix} \dfrac{\delta(1-\delta)}{N} & -\dfrac{\delta(\hat{p}_{11})_{p_{01}=\delta}}{N} \\ -\dfrac{\delta(\hat{p}_{11})_{p_{01}=\delta}}{N} & \dfrac{(\hat{p}_{11})_{p_{01}=\delta}[1-(\hat{p}_{11})_{p_{01}=\delta}]}{N} \end{pmatrix}, \tag{9}$$

So, with a large N,

$$z_a = \left(\frac{x_{01}}{N} - \delta \right) \Big/ \sqrt{\frac{\delta(1-\delta)}{N}} = \frac{x_{01} - N\delta}{\sqrt{N\delta(1-\delta)}} \tag{10}$$

will be approximately normally distributed with zero mean and unit variance.

Now we have established one statistical method of testing whether or not the detection rate of signals in one considered treatment is significantly inferior to that in the standard treatment that is theoretically superior to any other considered treatment in the detection of signals, i.e., one rejects H_0 in favor of H_1 when $z_a > z_\alpha$, where z_α is the point that cuts off $100\times\alpha$ percent of the area of the upper tail of the normal distribution with zero mean and unit variance; that is,

$$\frac{1}{\sqrt{2\pi}} \int_{z_a}^{\infty} \exp\left(-\frac{x^2}{2} \right) dx = \alpha . \tag{11}$$

Score method

By using the general theory of Bartlett[1, 2], we can derive another test statistic. This derivation is similar to the method given by Nam[3].

The first-order partial derivatives of the log likelihood, L, are

$$\partial L / \partial p_{01} = x_{01} / p_{01} - (N - x_{01} - x_{11}) / (1 - p_{01} - p_{11}) \quad (12)$$

and

$$\partial L / \partial p_{11} = x_{11} / p_{11} - (N - x_{01} - x_{11}) / (1 - p_{01} - p_{11}). \quad (13)$$

The MLE of p_{11} for a given value of $p_{01} = \delta$, $(\hat{p}_{11})_{p_{01} = \delta}$, has already been stated. According to Bartlett[2], let us consider the quantity

$$\frac{\partial L}{\partial p_{01}} - \frac{I_{12}}{I_{22}} \frac{\partial L}{\partial p_{11}} . \quad (14)$$

Its mean and variance are known as 0 and $I_{11} - I_{12}^2 / I_{22}$, and, in large samples, it can be considered to be normally distributed. From Eq. (6), $I_{11} - I_{12}^2 / I_{22}$ is given as,

$$I_{11} - \frac{I_{12}^2}{I_{22}} = \frac{N}{p_{01}(1 - p_{01})} . \quad (15)$$

Now we get the score statistics for testing the null hypothesis H_0: $p_{01} = \delta$ against the alternative H_1: $p_{01} > \delta$. This statistic is expressed as

$$z_s = \frac{x_{01} - N\delta - x_{01}(\hat{p}_{11})_{p_{01} = \delta} + x_{11}\delta}{(1 - \delta - (\hat{p}_{11})_{p_{01} = \delta})} \left(\frac{1 - \delta}{N\delta} \right)^{1/2} . \quad (16)$$

So, we reject H_0 in favor of H_1 when $z_s > z_\alpha$. Here, z_s is simplified as

$$z_s = \frac{x_{01} - N\delta}{\sqrt{N\delta\,(1 - \delta)}} . \quad (17)$$

Therefore, the score test statistic z_s is algebraically the same as the statistic from the maximum likelihood method z_a.

AN EXAMPLE

As an example of some applications of the above statistical method, let us consider the study described in the introduction. In this study, we investigated CRT monitor luminance in which the detection rate of nodules is reliably inferior to that in a CRT monitor with standard luminance. For the image reading study, we posited the 10 monitor conditions; the luminance level of these display monitors was lower than that of the standard display monitor.

For the image-reading experiment, 11 posteroanterior chest radiographs were acquired from normal volunteers. Five simulated nodules with different contrasts and diameters were digitally superimposed on these radiographs. Thirteen radiologists observed the 11 chest X-ray soft-copy images under the above-mentioned ten darker monitor display conditions and the standard monitor condition, and tried to detect the nodules on each image (each image included the five artificial nodules). Here, for each image, the observation order was always from the darkest display condition to the brightest. For each observer, we calculated the rate of detecting the 55 nodules correctly for each monitor display condition. Our aim was to determine the display conditions under which the detection rate of nodules would show statistically significant inferiority compared with the standard display condition.

In this study, the number of nodules detected in the display monitors with the darker luminance was always smaller than the one in the standard display monitor. Therefore, the statistical analysis for this experiment is a good example of our method.

Adopting 5% as a significance level for the statistical test, we will conduct 10 successive tests for each observer. Thus, we adopt the cut off point of 5/10 = 0.5 percent of the area of

the upper tail of the normal distribution with zero mean and unit variance for z_α, even though it is the most conservative. Then we reject H_0 in favor of H_1 when $z_a > 2.5758$ or $z_s > 2.5758$. Furthermore, we adopt 0.01 as a value of δ.

Table 2 shows the results of one observer, who detected 45 nodules on the monitor with the standard luminance. The test statistics were calculated under the condition of $\delta = 0.01$. The higher the monitor condition number was, the brighter the luminance of the CRT monitor.

DISCUSSION

In this way we arrived at the statistical method in a comparative study using a matched-pairs design in which the standard treatment is theoretically or practically superior to the others. For such a problem, one may use McNemar's test. However, since in the case of $x_{10} = 0$, the standard McNemar's test cannot be used, one must modify it. Our method is one of these modifications, enabling one to use a very sensitive statistical method for such a problem.

Compared with the method of using the MLE directly, the score method is theoretically superior[1]. As is obvious from Eq. (16), the statistic from the score method is superficially complex and difficult to understand. However, as we have shown, the score test statistic is algebraically the same as the statistic derived directly from the maximum likelihood method. The equivalence between these two statistics is consistent with the one between the maximum likelihood test statistic and the score test statistic that was shown by Nam[3].

In our method, the selection of a value of δ is arbitrary. Usually, this quantity should be equivalent to the measurement error in the considered experiment. However, as in the example given in this paper, this measurement error is not obvious. In our experiment establishing the threshold luminance level of electronic displays for medical image interpretation, we think that, for clinical use, the number of detected nodules under considered monitor conditions should be equal to that in the standard monitor; that is, if the number of detected nodules in one monitor condition decreases, even by one, compared to that in the standard monitor, such a monitor condition should be considered inferior to that in the standard monitor. In our image reading experiment, the total number of nodules that each observer had to detect is 55. Therefore, we

Table 2. Results of one observer in the image reading test. This observer detected 45 nodules in the standard monitor. Test statistics were calculated under the condition of $\delta = 0.01$. The higher the number of monitor conditions, the brighter the luminance of the CRT monitor.

Monitor condition	Detected nodules	Test statistic	
1	11	45.331196	significant
2	17	37.200043	significant
3	24	27.713690	significant
4	28	22.292919	significant
5	33	15.516957	significant
6	36	11.451379	significant
7	40	6.030608	significant
8	41	4.675415	significant
9	43	1.965030	not significant
10	43	1.965030	not significant

think that the difference in the detection rates between the considered treatment and the standard treatment should be below $1/55 \approx 0.018$, and we adopt 0.01 as a value of δ. However, opinions may differ on this point.

In the statistical tests used in our example, we adopted the most conservative method among multiple comparison procedures. Further consideration should be given to developing more sophisticated methods on this subject.

REFERENCES

1) Bartlett, M.S.: Approximate confidence intervals. *Biometrika*, 40, 12–19 (1953).
2) Bartlett, M.S.: Approximate confidence intervals II. more than one unknown parameter. *Biometrika*, 40, 306–317 (1953).
3) Nam, J.M.: Establishing equivalence of two treatments and sample size requirements in matched-pairs design. *Biometrics*, 53, 1422–1430 (1997).